IGGSA Shared Tasks on German Sentiment Analysis (GESTALT)

Josef Ruppenhofer[‡], **Roman Klinger**[*†], **Julia Maria Struß**[‡],
Jonathan Sonntag[§], **Michael Wiegand**[◦]
[‡] Dept. of Information Science and Language Technology, Hildesheim University
[†] Institute for Natural Language Processing, University of Stuttgart
[*] Semantic Computing Group, CIT-EC, Bielefeld University
[§] Computational Linguistics, Potsdam University
[◦] Spoken Language Systems, Saarland University
{ruppenho,julia.struss}@uni-hildesheim.de
roman.klinger@ims.uni-stuttgart.de
jonathan.sonntag@yahoo.de
michael.wiegand@lsv.uni-saarland.de

Abstract

We present the German Sentiment Analysis Shared Task (GESTALT) which consists of two main tasks: *Source, Subjective Expression and Target Extraction from Political Speeches (STEPS)* and *Subjective Phrase and Aspect Extraction from Product Reviews (StAR)*. Both tasks focused on fine-grained sentiment analysis, extracting aspects and targets with their associated subjective expressions in the German language. STEPS focused on political discussions from a corpus of speeches in the Swiss parliament. StAR fostered the analysis of product reviews as they are available from the website Amazon.de. Each shared task led to one participating submission, providing baselines for future editions of this task and highlighting specific challenges. The shared task homepage can be found at https://sites.google.com/site/iggsasharedtask/.

1 Introduction

In opinion mining, we are not only interested in detecting the presence of opinions (or more broadly, subjectivity) but determining particular attributes. We want to determine *which* valence or polarity an opinion has (positive, negative or neutral), *how* strong it is (intensity), and also know *whose* opinion it is and *what* it is about. The last two questions are what the task of opinion source and target extraction is concerned with. Source and target extraction are capabilities needed for the analysis of unrestricted language texts, where this kind of information cannot be derived from meta-data and where opinions by multiple sources and about multiple, potentially related, targets appear side by side.

We present two shared tasks that ran under the auspices of the Interest Group of German Sentiment Analysis[1] (IGGSA). Maintask 1 on *Source, Subjective Expression and Target Extraction from Political Speeches (STEPS)* constitutes the first evaluation campaign for source and target extraction on German language data. Maintask 2 on *Subjective Phrase and Aspect Extraction from Product Reviews (StAR)* focuses on the aspect extraction, which is understood as the target of a subjective phrase. For both tasks, publicly available resources have been created, which serve as a reference corpus for the evaluation of opinion source and target extraction in German.

2 Task Descriptions

In this section, we present the task setting, describe the dataset, the annotation, the subtasks, the evaluation and results for each of the two main tasks (Section 2.1 and Section 2.2), respectively.

2.1 Maintask 1

Maintask 1 calls for the identification of subjective expressions, sources and targets in parliamentary speeches. While these texts can be expected to be opinionated, they pose the challenges that

[1] https://sites.google.com/site/iggsahome/

sources other than the speaker may be relevant and that the targets, though constrained by topic, can vary widely. As in the case of Maintask 2, the dataset provided is the first one that provides publicly available expression-level annotations on running texts of this type for German.

2.1.1 Dataset

The STEPS data set stems from the debates of the Swiss parliament (*Schweizer Bundesversammlung*).[2] This particular data set was selected for two reasons. First, the source data is open to the public and we can re-distribute it with our annotations. We were not able to fully ascertain the copyright situation for German parliamentary speeches, which we had also considered. Second, the text calls for annotation of multiple sources and targets.

As the Swiss parliament is a multi-lingual institution, we were careful to exclude not only non-German speeches but also German speeches that constitute responses to, or comments on, speeches, heckling, and side questions in other languages. This way, our annotators did not have to label any German data whose correct understanding might rely on material in a language that they might not be able to interpret correctly.

Some potential linguistic difficulties consisted in peculiarities of Swiss German found in the data. For instance, the vocabulary of Swiss German is different from standard German, often in subtle ways. For instance, the verb *vorprellen* is used in the following example instead of *vorpreschen*, which would be expected for German spoken in Germany:

> *Es ist unglaublich: Weil die Aussenministerin vorgeprellt ist, kann man das nicht mehr zurücknehmen. (Hans Fehr, Frühjahrsession 2008, Zweite Sitzung – 04.03.2008)*[3]
>
> 'It is incredible: because the foreign secretary acted rashly, we cannot take that back again.'

In order to reduce any negative impact that might come from misreadings of the Swiss German by our annotators, who were German and Austrian rather than Swiss, we selected speeches about what we deemed to be non-parochial issues. For instance, we picked texts on international affairs rather than ones about Swiss municipal governance.

Technically, the STEPS data underwent the following pre-processing pipeline. Sentence segmentation and tokenization was done using OpenNLP[4], followed by lemmatization with the TreeTagger (Schmid, 1994), constituency parsing by the Berkeley parser (Petrov and Klein, 2007), and final conversion of the parse trees into TigerXML-Format using TIGER-tools (Lezius, 2002). To perform the annotation we used the Salto-Tool (Burchardt et al., 2006).

2.1.2 Annotation

Through our annotation scheme[5], we provide annotations at the expression level. No sentence or document-level annotations are manually performed or automatically derived.

There were no restrictions imposed on annotations. The subjective expressions could be verbs, nouns, adjectives or multi-words. The sources and targets could refer to any actor or issue as we did not focus on anything in particular.

The definition of subjective expressions (SE) that we used is broad and based on well-known prototypes. It largely follows the model of what Wilson and Wiebe (2005) subsume under the umbrella term *private state*, as defined by Quirk et al. (1985): "As a result, the annotation scheme is centered on the notion of private state, a general term that covers opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments.":

- evaluation (positive or negative): *toll* 'great', *doof* 'stupid'

[2] The full task test data is available at `https://sites.google.com/site/iggsasharedtask/home/testdata-maintask1-salto_tiger-xml.zip`. The subtask test data for is at `https://sites.google.com/site/iggsasharedtask/home/testdata-maintask1-subtasks-salto_tiger.xml.zip`.

[3] `http://www.parlament.ch/ab/frameset/d/n/4802/263473/d_n_4802_263473_263632.htm`

[4] `http://opennlp.apache.org/`

[5] See `https://sites.google.com/site/iggsasharedtask/task-1/STEPS_guide.pdf` for the the guidelines we used.

Name	Source		Target		Frame	
SwissGerman	*not applicable*				14	
RhetoricalDevices	*not applicable*				64	
Inferred	344	(7.8%)	177	(3.9%)	97	(2.0%)
Uncertain	61	(1.4%)	29	(0.6%)	58	(1.2%)

Table 1: Flags annotated across all annotators and files of Maintask 1

	F_1	Dice for true positives
Subjective Expression	63.32	0.92
Sources*	68.70	0.99
Targets*	80.63	0.85

Table 2: Average inter-annotator agreement across all pairs of annotators on test data of Maintask 1 (F_1 is based on partial overlap; Dice quantifies the amount of overlap for matches)

- (un)certainty:
  zweifeln 'doubt', *gewiss* 'certain'
- emphasis:
  sicherlich/bestimmt 'certainly'
- speech acts:
  sagen 'say', *ankündigen* 'announce'
- mental processes:
  denken 'think', *glauben* 'believe'

Beyond giving the prototypes, we did not seek to impose on our annotators any particular definition of subjective or opinion expressions from the linguistic, natural language processing or psychological literature related to subjectivity, appraisal, emotion or related notions.

In marking subjective expressions, the annotators were told to select minimal spans. This guidance was given because we had decided that within the scope of this shared task we would forgo any treatment of polarity and intensity. Accordingly, negation, intensifiers and attenuators and any other expressions that might affect a minimal expression's polarity or intensity could be ignored.

When labeling sources and targets, annotators were asked to first consider syntactic and semantic dependents of the subjective expressions. If sources and targets were locally unrealized, the annotators could annotate other phrases in the context. Where a subjective expression represented the view of the implicit speaker or text author, annotators could indicate this by setting a flag *Sprecher* 'Speaker' on the the source element.

For all three types of labels, subjective expressions, sources, and targets, annotators had the option of using two additional flags. The first flag was intended to mark a label instance as *Inferiert* 'Inferred'. In the case of subjective expressions, this covers, for instance, cases where annotators were not sure if an expression constituted a polar fact or an inherently subjective expression. In the case of sources and targets, the 'inferred' label applies to cases where the referents cannot be annotated as local dependents but have to be found in the context. The second flag afforded annotators the ability to mark an annotation as *Unsicher* 'Uncertain', if they were unsure whether the span should really be labeled with the relevant category.

The annotators were asked to use a flag *Rhetorisches Stilmittel* 'Rhetorical device' for subjective expression instances where subjectivity was conveyed through some kind of rhetorical device such as repetition. Across all three annnotators, 64 instances were labeled as 'rhetorical de-

Run	Measure	Subjective Expression	Source	Source_SE	Target	Target_SE
Run 3	Prec	63.42	**48.55**	**74.89**	**56.25**	79.71
	Rec	26.10	11.32	**42.46**	15.60	**58.00**
	F_1	36.98	18.36	**54.19**	24.43	**67.14**
Run 5	Prec	**80.56**	47.98	58.55	*not applicable*	
	Rec	29.97	10.44	32.65	*not applicable*	
	F_1	43.69	17.14	41.92	*not applicable*	

Table 3: Best participant runs for Maintask 1 (3 = rule-based system; 5 = translation-based system, which did not include Targer identification. Results suffixed with subjective expressions consider only cases where the system already matched the gold standard on the subjective expression)

vice' in the data.

Finally, the annotation guidelines gave annotators the option to mark particular subjective expressions as *Schweizerdeutsch* 'Swiss German' when they involved language usage that they were not fully familiar with. Such cases could then be excluded or weighted differently for the purposes of system evaluation. In our annotation, these markings were in fact rare with only 14 of such flag instances across all three annotators.

Summing over all three annotators, our dataset covers 1815 sentences. In total, 4935 subjective expression frame instances were labeled by the annotators combined (2.7 frames/sentence). Related to the frames, 8959 frame element (source or target) instances were annotated (1.8 frame elements/frame). Although the theory embodied by our guidelines calls for at least one source and target label per annotated subjective expression frame, we find slightly less than one instance of each (4427 sources, 4532 targets). In Table 1, we see that not many flags were annotated by our annotators. The careful selection of our data with respect to the topics treated seems to have worked well. We have few instances of subjective expressions that were flagged as Swiss German formulations by our annotators. The most common type of flag was the one for 'inferred' labels. Here, inference of sources was by far the most common case. Note, that fewer labels were marked 'uncertain' than were marked 'inferred'. Inference did not necessarily result in uncertainty.

In Table 2, we present results on the inter-annotator agreement on the test data. One way of measuring the agreement uses the precision/recall-framework of evaluation. We calculate the relevant numbers based on treating one annotator as gold and another as system, and averaging the results for the three pairs of annotators. For F_1, we counted a true positive when there was partial span overlap. In addition, we present a token-based multi-κ value (Davies and Fleiss, 1982). Given that in our annotation scheme, a single token can be e.g. a target of one subjective expression while itself being a subjective expression as well, we need to calculate three kappa values covering the binary distinctions between presence of each label and its absence. For subjective expressions κ is 0.39, for sources 0.57, and for targets 0.46.

As exact matches on spans are relatively rare, the Dice coefficient is used to measure the overlap between a system annotation and a gold standard annotation (Dice, 1945). The Dice coefficient $\mathrm{dc}(S, G)$ is a similarity measure ranging from 0 to 1, where

$$\mathrm{dc}(S, G) = \frac{2|S \cap G|}{|S| + |G|},$$

and G is the set of tokens in the gold annotations and S the set of tokens the prediction (the system label), respectively.

2.1.3 Subtasks

The STEPS shared task offered a full task as well as two subtasks:

Full task Identification of subjective expressions with their respective sources and targets.

Subtask 1 Participants are given the subjective expressions and are only asked to identify opinion sources.

Subtask 2 Participants are given the subjective expressions and are only asked to identify opinion targets.

Participants could choose any combination of the tasks. However, so as to not give an unfair advantage, the full task was run and evaluated before the gold information on subjective expressions was given out for the two subtasks, which were run concurrently.

2.1.4 Evaluation Metrics

The runs that were submitted by the participants of the shared task were evaluated on different levels, according to the task they chose to participate in. For the full task, there was an evaluation of the subjective expressions as well as the targets and sources for subjective expressions, matching the system's annotations against those in the gold standard. For subtasks 1 and 2, only the sources and targets were evaluated, as the subjective expressions were already given.

In this first iteration of the STEPS task, we evaluated against each of our three annotators individually rather than against a single gold-standard. Our intent behind this choice was to retain the variation between the annotators.

We used recall to measure the proportion of correct system annotations with respect to the gold standard annotations. Additionally, precision was calculated so as to give the fraction of correct system annotations relative to all the system annotations. As we did for inter-annotator-agreement, for recall and precision we counted a match when there was partial span overlap. Similarly, we again used the Dice coefficient to assess the overlap between a system annotation and a gold standard annotation.

The group that participated in our main task submitted five different runs, based on two different system architectures. Table 3 shows the best result for each architecture. The scores represent averages across the comparisons relative to each of the three annotators. The rule-based system generally performed better than the translation-based one. However, the latter was much better in its precision on recognizing subjective expressions in the full task. As is to be expected, when the system had already matched the gold standard on the subjective expressions, its performance on source and target recognition, shown in columns Source_SE , Target_SE, is much superior to performance in the general case.

2.2 Maintask 2: Subjective Phrase and Aspect Extraction from Product Reviews

Maintask 2 was designed to foster the development of systems to automatically extract subjective, evaluative phrases from German Amazon reviews, aspects described in the review and their relation, i.e., which evaluative phrase targets which aspect. In addition, another focus is cross-domain learning: The development corpus consists of reviews for various products while the test corpus is from yet another product not known to the participants before.

2.2.1 Dataset

For this task, a data set was provided for training parameters and developing the system. *The USAGE Review Corpus for Fine Grained Multi Lingual Opinion Analysis* (Klinger and Cimiano, 2014) was previously published and was fully available to the participants from the start of the task on. It consists of 611 German and 622 English reviews for coffee machines, cutlery sets, microwaves, toasters, trashcans, vacuum cleaners, and washers from which only the German part has been used in this shared task. To construct the test corpus, 1646 reviews for the search term *Wasserkocher* 'water boiler' were retrieved. From these, 100 sampled reviews were annotated and included in the test corpus. The training[6] and test[7] data is freely available.

2.2.2 Annotation

The entity classes *aspect* and *evaluative (subjective) expression* are annotated in the corpus. Evaluative expressions are assigned a polarity (posi-

[6]Maintask 2 training data: `http://dx.doi.org/10.4119/unibi/citec.2014.14`

[7]Maintask 2 test data: `http://dx.doi.org/10.4119/unibi/2695161`

tive, negative, neutral), which is not used in this shared task, and a set of aspects they refer to. The annotators were instructed to regard everything as an aspect that is part of a product or related to it and can influence the opinion about it, including the whole product itself. Evaluative phrases express an opinion. Negations are not separately annotated but are part of a phrase. Annotators were asked to avoid overlapping annotations if possible. The annotations should be as short as possible, as long as the meaning is understandable if only the annotations were given (without the sentence itself).

Every review in the training data is annotated by two linguists, the test data is annotated by one (the information which of the training data annotation corresponds to the annotator of the test data is available).

In the following examples, aspects are marked in blue and subjective phrases are marked in red:

Ich hatte keine Probleme mit der Rückgabe .
I had no problems with the return .

return is a target of *no problems*.
no problems is positive.

Die Waschmaschine selbst ist toll , der beiliegende Schlauch ist Schrott.
The washer itself is great , the included hose is junk .

washer is a target of *great*.
hose is a target of *junk*.
great is positive.
junk is negative.

Es sieht sehr hübsch aus, wie ein Aufbewahrungsbehälter , er ist leicht und einfach zu benutzen .
It looks very neat , like a storage container , and using it is very simple and easy .

– *looks* is a target of *very neat*.
using is a target of *simple* and of *easy*.

The inter-annotator agreement of the full training corpus is $\kappa = 0.65$ (Cohen's κ). The inter-annotator F_1 measure is 0.71 for aspects, 0.55 for subjective phrases and 0.42 for the relations between both (including an error propagation of having the exact same phrases annotated). These measures can be regarded as upper bounds for meaningful results of an automated approach.

Table 4 presents the main statistics of the training and testing corpora. Here, annotator 1 of the training corpus performed the annotation of the test data. Obviously, the number of annotated phrases is higher in the test data.

The most frequent subjective phrases for the different products are very similar. For instance, the phrases *gut* 'good' and *sehr zufrieden* 'very satisfied' occurs in all top 10 lists of subjective phrases. However, the most frequent aspect phrases are very different, as the product category itself is frequently used as an aspect (e.g. *Kaffeemaschine* 'coffee maker' or *Besteck* 'cutlery'). In addition, very product class-specific aspects are mentioned frequently, like *Wasser* 'water', *schneiden* 'cut', or *Edelstahl* 'stainless steel'. Some aspects are shared between product categories, for instance *Preis* 'price' or *Qualität* 'quality'.

Clearly, the cross-domain inference task is more challenging, as the mentioned aspects are not as similar as the annotated subjective phrases.

2.2.3 Subtasks

The three substasks to be addressed by the participants were:

Subtask 2a Identication of subjective phrases.

Subtask 2b Identification of aspect phrases.

Subtask 2c Identification of subjective phrases and aspect phrases and indication for each aspect phrase of which subjective phrase it is the target (if any).

2.2.4 Evaluation metrics and Baseline approach

For evaluation, the F_1 measure of the exact match of the predicted phrases in comparison to the annotated phrases is taken into account. This is straight-forward for Subtasks 2a and 2b. In 2c, a pair of aspect and subjective phrase was considered to be correctly identified, if both phrases

	Train Ann. 1	Train Ann. 2	Test
Number of reviews	611		100
Number of products	127		100
Number of Aspects	6340	5055	1662
Number of Aspects/Review	10.4	8.3	16.6
Number of positive Subj.	3840	3717	823
Number of positive Subj./Review	6.3	6.1	8.2
Number of negative Subj.	1094	1052	264
Number of negative Subj./Review	1.8	1.7	2.6
Target Rel.	4085	4643	1013
Target Rel./Review	6.7	7.6	10.1

Table 4: Statistics of the corpora used in Maintask 2

predicted to be participating were identified correctly (on the phrase level) as well as annotated as a pair.

For comparison, as a baseline, a machine learning-based system optimized for in-domain inference was applied[8] (Klinger and Cimiano, 2013a; Klinger and Cimiano, 2013b). A comparison of the participant's result and the baseline is shown in Table 5. It can be observed that the baseline outperforms the subjective phrase detection, but the result submitted by the participant is superior in the more difficult cross-domain tasks of aspect extraction. The extraction of relations clearly remains a challenge.

3 Related Work

While quite a few shared tasks have addressed the recognition of subjective units of language and, possibly, the classification of their polarity (SemEval 2013 Task 2, Twitter Sentiment Analysis (Nakov et al., 2013); SemEval-2010 task 18: Disambiguating sentiment ambiguous adjectives (Wu and Jin, 2010); SemEval-2007 Task 14: Affective Text (Strapparava and Mihalcea, 2007) *inter alia*), few tasks have included the extraction of sources and targets.

The prior work most relevant to the tasks presented here was done in the context of the Japanese NTCIR[9] Project. In the NTCIR-6 Opinion Analysis Pilot Task (Seki et al., 2007), which was offered for Chinese, Japanese and English, sources and targets had to be found relative to whole opinionated sentences rather than individual subjective expressions. However, the task allowed for multiple opinion sources to be recorded for a given sentence if there were multiple expressions of opinion. The opinion source for a sentence could occur anywhere in the document. In the evaluation, as necessary, co-reference information was used to (manually) check whether a system response was part of the correct chain of co-referring mentions. The sentences in the document were judged as either relevant or non-relevant to the topic (=target). Polarity was determined at the sentence level. For sentences with more than one opinion expressed, the polarity of the main opinion was carried over to the sentence as a whole. All sentences were annotated by three raters, allowing for strict and lenient (by majority vote) evaluation. The subsequent Multilingual Opinion Analysis tasks NTCIR-7 (Seki et al., 2008) and NTCIR-8 (Seki et al., 2010) were basically similar in their setup to NTCIR-6.

While GESTALT shared tasks focussed on German, the most important difference to the shared tasks organized by NTCIR is that it defined the source and target extraction task at the level of individual subjective expressions. There was no comparable shared task annotating at the expression level, rendering existing guidelines imprac-

[8] A high-recall combination of the joint configuration and the pipeline setting has been applied.

[9] NII [National Institute of Informatics] Test Collection for IR Systems

	Baseline			Participant		
Subtask	Prec	Rec	F_1	Prec	Rec	F_1
Aspect Phrase	**65.5**	46.4	54.3	55.5	62.2	**58.7**
Subjective Phrase	51.5	**41.4**	**45.9**	**51.6**	32.0	39.5
Relation	**15.9**	8.3	10.9	12.6	**13.8**	**13.2**

Table 5: Results of the baseline system and the participant's best submission in Maintask 2.

tical and necessitating the development of completely new guidelines.

Another more recent shared task related to GESTALT is the Sentiment Slot Filling track (SSF) that was part of the Shared Task for Knowledge Base Population of the Text Analysis Conference (TAC) organised by the National Institute of Standards and Technology (NIST) (Mitchell, 2013). The major distinguishing characteristic of that shared task, which is offered exclusively for English language data, lies in its retrieval-like setup. Here, the task is to extract all possible opinion sources and targets from a given text. By contrast, in SSF the task is to retrieve sources that have some opinion towards a given target entity or targets of some given opinion sources. In both cases, the polarity of the underlying opinion is also specified within SSF. The given targets or sources are considered a type of *query*. The opinion sources and targets are to be retrieved from a document collection.[10] Unlike GESTALT, SSF uses heterogeneous text documents including both newswire and discussion forum data from the Web.

This year's SemEval-2014 Task 4 on Aspect Based Sentiment Analysis (ABSA) on English review data for restaurant and laptop reviews (Pontiki et al., 2014) constitutes another related shared task. It focused on aspect-based polarity detection. The main differences are that the aspect categories were predefined and that the polarity assignment did not include the detection of the evaluative phrases. Therefore, the polarity assignment was on the aspect level and the relation between a subjectivity-bearing word was implicit. Another difference between ABSA and GESTALT (StAR, specifically) is that the number of products taken into account is higher in StAR, motivating a cross-domain inference challenge.

[10] In 2014, the text from which entities are to be retrieved is restricted to one document per query.

4 Conclusion and Outlook

We reported on the first iteration of two shared tasks for German sentiment analysis. Both tasks focused on the discovery of subjective expressions and their related entities. In the case of STEPS, sources and targets had to be found and linked to subjective expressions in political speeches, in the case of StAR, aspects had to be identified and tied to subjective expressions in Amazon reviews.

Although a preliminary call for interest had indicated interest by 3–4 groups for each of the tasks, in the end each task had only one participant. We therefore solicited feedback from actual and potential participants at the end of the IGGSA-GESTALT workshop in order to be able to tailor the tasks better in a future iteration.

Based on the discussion, both shared tasks plan on including polarity in the evaluation for their next iteration. For both tasks, there was discussion what a suitable evaluation procedure would be, in particular whether partial matches should be the basis of the main measures or if exact matches would be more desirable.

Specific to STEPS, we are considering conducting the evaluation in alternative ways on a future iteration of the task. One direction to pursue is to derive new versions of the gold standard based on the level of inter-annotator agreement on the labels. In a full-agreement mode, we would only retain annotations of the gold standard that had majority or even full agreement on the subjective expression level for all three annotators. Another alternative would consist in establishing an expert-adjudicated gold-standard, after all. The benefit of any of these alterna-

tive evaluation modes would be that a clear objective function can be learnt and that the upper bound for system performance would again be 100% precision/recall/F_1-score, whereas it was lower for this iteration given that existing differences between the annotators necessarily led to false positives and negatives.

For the next iteration of GESTALT, we plan to make a baseline system available, such that the barrier to participation in the shared task is lower and participants' efforts can be focused on the actual methods.

Acknowledgments

We would like to thank Simon Clematide for helping us get access to the Swiss data for the STEPS task. For their support in preparing and carrying out the annotations of this data, we would like to thank Jasper Brandes, Melanie Dick, Inga Hannemann, and Daniela Schneevogt. We thank the German Society for Computational Linguistics for its financial support of the STEPS annotation effort. For the annotations used in the StAR task, we thank Luci Fillinger and Frederike Strunz. Roman Klinger was partially funded by the *It's OWL* project ('Intelligent Technical Systems Ostwestfalen-Lippe', `http://www.its-owl.de/`), a leading-edge technology and research cluster funded by German Ministry of Education and Research (BMBF). This first and last author were partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1, respectively.

References

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, and Sebastian Pado. 2006. SALTO - A Versatile Multi-Level Annotation Tool. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 517–520.

Mark Davies and Joseph L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, 38(4):1047–1051.

Lee R. Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

Roman Klinger and Philipp Cimiano. 2013a. Bidirectional inter-dependencies of subjective expressions and targets and their value for a joint model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 848–854, Sofia, Bulgaria, August. Association for Computational Linguistics.

Roman Klinger and Philipp Cimiano. 2013b. Joint and pipeline probabilistic models for fine-grained sentiment analysis: Extracting aspects, subjective phrases and their relations. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pages 937–944, Dec.

Roman Klinger and Philipp Cimiano. 2014. The usage review corpus for fine grained multi lingual opinion analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Wolfgang Lezius. 2002. TIGERsearch - Ein Suchwerkzeug für Baumbanken. In Stephan Busemann, editor, *Proceedings of KONVENS 2002*, Saarbrücken, Germany.

Margaret Mitchell. 2013. Overview of the TAC2013 Knowledge Base Population Evaluation: English Sentiment Slot Filling. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, MD, USA.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta and Georgia and USA. Association for Computational Linguistics.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Randolph Quirk, Sidney Greenbaum, Geoffry Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Longman.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Hsin-Hsi. Chen, Noriko Kando, and Chin-Yew Lin. 2007. Overview of opinion analysis pilot task at ntcir-6. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 265–278.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando. 2008. Overview of multilingual opinion analysis task at NTCIR-7. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, pages 185–203.

Yohei Seki, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando. 2010. Overview of Multilingual Opinion Analysis Task at NTCIR-8: A Step Toward Cross Lingual Opinion Analysis. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 209–220.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74. Association for Computational Linguistics.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 53–60, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Yunfang Wu and Peng Jin. 2010. SemEval-2010 Task 18: Disambiguating Sentiment Ambiguous Adjectives. In Katrin Erk and Carlo Strapparava, editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 81–85, Stroudsburg and PA and USA. Association for Computational Linguistics.